Using sketch-map coordinates to analyze and bias molecular dynamics simulations

Tribello, G. A., Ceriotti, M., & Parrinello, M. (2012). Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(14), 5196-5201. https://doi.org/10.1073/pnas.1201152109

Published in:
Proceedings of the National Academy of Sciences of the United States of America

Document Version:
Peer reviewed version

Using sketch-map coordinates to analyze and bias molecular dynamics simulations

Gareth A. Tribello * Michele Ceriotti † and Michele Parrinello *

*Computational Science, Department of Chemistry and Applied Biosciences, ETHZ Zurich USI-Campus, Via Giuseppe Buffi 13 C-6900 Lugano, and †Physical and Theoretical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, United Kingdom

Submitted to Proceedings of the National Academy of Sciences of the United States of America

When examining complex problems, such as the folding of proteins, coarse grained descriptions of the system drive our investigation and help us to rationalize the results. Oftentimes collective variables (CVs), derived through some chemical intuition about the process of interest, serve this purpose. So, because finding these CVs is the most difficult part of any investigation, we recently developed a dimensionality reduction algorithm, sketch-map, that can be used to build a low-dimensional map of a phase space of high-dimensionality. In this paper we discuss how these machine-generated CVs can be used to accelerate the exploration of phase space and to reconstruct free-energy landscapes. To do so, we develop a formalism in which high-dimensional configurations are no longer represented by low-dimensional position vectors. Instead, for each configuration we calculate a probability distribution which has a domain that encompasses the entirety of the low-dimensional space. To construct a biasing potential we exploit an analogy with metadynamics and use the trajectory to adaptively construct a repulsive, history-dependent bias from the distributions that correspond to the previously-visited configurations. This potential forces the system to explore more of phase space by making it desirable to adopt configurations whose distributions do not overlap with the bias. We apply this algorithm to a small model protein and succeed in reproducing the free energy surface that we obtain from a parallel tempering calculation.

dimensionality reduction | free energy | enhanced sampling | proteins

Introduction

Statistical mechanics connects the micro and macro scales by showing how thermodynamic state functions, such as the free energy, can be calculated from the classical Hamiltonians that govern the motions of atoms and molecules. These equations allow us to calculate ensemble averages, the relative stabilities of structures and in some cases reaction mechanisms. At first glance the $3N$-dimensional integrals over configuration space make the equations of statistical mechanics appear unsolvable. However, all of them involve integrals over distributions in which the probability of a microstate is related to its energy. Therefore, because the vast majority of phase space is energetically inaccessible, only a relatively small number of configurations make non-negligible contributions (1–4). Hence, the problem is not so much the integrals - rather it is determining which are the low energy states that significantly contribute to them.

Molecular dynamics (MD) - using Newton's equations to calculate a trajectory for the system - is a technique that we can use to find the energetically-accessible portions of phase space. The configurations visited during an MD simulation are distributed according to the canonical ensemble so ensemble averages can be calculated by just averaging over the trajectory. However, in doing this one has to assume ergodicity, i.e. that *all* relevant configurations have been visited during the simulation. This is a problem whenever the energy landscape contains long-lived stable/metastable minima separated by high barriers (5). These features dramatically decrease the rate at which phase space is sampled and so introduce a characteristic time scale for phenomena, which, for protein folding and phase transitions, is typically on the order of milliseconds or more. Studying these process using unbiased MD is therefore difficult as when using this technique it is only possible to simulate the system for very short ($\approx$1 μs) periods of time. Admittedly, this limit can be extended (to $\approx$1 ms) by using specialized hardware but in doing this one is forced to limit the form of the Hamiltonian (6).

It is possible to increase the frequency with which the barriers separating metastable basins are crossed by introducing a bias potential that makes the energies in the basins comparable with the energies at the transition states (7). Furthermore, because we know the form of the bias function, we can reweight the biased trajectory and thereby obtain the unbiased free energy surface (8–11). These so called enhanced sampling methods are now commonplace and applying them to simple chemical problems is relatively straightforward (12–14). The problem comes when the chemistry is more complex, in large part because it is then not obvious how to construct the biasing potential using chemical/physical intuition.

Bias potentials are typically constructed as a function of a small number of collective variables (CVs). Selecting these CVs is the most difficult part of any investigation so we have recently begun to develop an automated strategy based on machine learning. The first step in this strategy is to obtain a very thorough sampling of the accessible portion of phase space using an algorithm, which adaptively constructs a bias as a function of a large number (D) of collective variables (15). By applying dimensionality reduction - in particular our recently developed sketch-map algorithm (16) - to the trajectory obtained from this calculation, one can obtain a lower, d-dimensional, representation of the accessible portion of phase space. Herein we present the final step of the process in which we adaptively construct a bias potential as a function of the sketch-map coordinates and thereby obtain a thorough sampling of phase space from which we can extract free energies through re-weighting. In what follows we first present the mathematical concepts and demonstrate the application of the algorithm on a simple model potential. We then apply it to the alanine 12 system that we examined in our two previous articles (15, 16) and show that we can use our new metadynamics algorithm to reproduce the free energy surface obtained via parallel tempering.

Reserved for Publication Footnotes

Background

In all chemical systems the shape of the potential energy surface makes large portions of phase space inaccessible by placing energetic constraints on the geometry of the system (5). In many of the commonly used biasing methods we assume that this accessible portion of phase space lies on a low-dimensionality manifold that is embedded in the full dimensionality space. For many methods vectors (CVs) that describe this manifold are selected through chemical/physical intuition. However, this process of finding appropriate CVs is often far from straightforward (17) and so there is a strong temptation to look to see whether an automated process can be devised.

An ideal CV for biased dynamics should produce a map of phase space in which all the significant basins in the free energy surface are well separated. In addition the CVs should be constructed so that, during the biased dynamics, the system will be pushed along the lowest-lying transition pathways. Dimensionality reduction and manifold learning algorithms are tools that, at least in theory, allow us to develop such CVs. These algorithms construct a d-dimensional representation of a set of data points distributed in a D-dimensional space. This is done by projecting points in the low-dimensionality space in a way that reproduces the pairwise distances between the points in the high-dimensionality space. In the high dimensionality space, these pairwise distances can be the Pythagorean distance (multidimensional scaling) (18), the geodesic distance (isomap) (19, 20), a non-linear transformation of the Pythagorean distance (kernel PCA) (21, 22) or the diffusion distance (diffusion maps) (23–27). In contrast, when one endeavors to distribute points in the low dimensionality space, this is typically done so that it is the Pythagorean distances that reproduce the high-dimensionality distances. Additionally, in the vast majority of applications, this process of distance matching is not done by iteratively minimizing the discrepancies between the distances in the high and low dimensional spaces. Instead some algebra is performed on the matrix of D-dimensional distances which makes the optimization process deterministic (18).

The problem with these methods is that it is difficult to come up with general D-dimensional metrics that will by necessity produce a set of distances that can be reproduced in a low-dimensional, linear space (28, 29). As an example of how problematic this can be consider mapping the surface of a sphere in two dimensions, as one has to do to draw a map of the world. The resulting representation will inevitably provide a distorted view of the original. Furthermore, discontinuities can only be avoided if one incorporates a non-linear feature - the periodicity - in the low-dimensional representation. Worse still, and more relevant to the problem at hand, is the fact that in our previous paper we provided evidence that certain features in typical trajectory data are characteristic of a distribution of points in the full dimensionality space (16). These realizations led us to develop a new algorithm - sketch-map - for performing dimensionality reduction on trajectory data. In developing this algorithm we imagine that the free energy surface is composed of a network of energetic basins, connected by a spider's web of narrow transition pathways. Points distributed on this surface therefore display high-dimensionality features because the fluctuations within each basin take place in the full dimensional space, and because the basins are scattered across the D dimensions. Thus, in sketch-map we try to qualitatively reproduce the spiders web of connections by transforming the distances in both the D-dimensional and d-dimensional spaces. This transformation ensures that the algorithm focuses on reproducing the distances that lie within a particular range - the length scale that corresponds to the transition pathways between basins. For the remainder of the distances we only insist that if the points are close together in the high dimensionality space they should be projected close together, while if they are far apart they should be projected far apart.

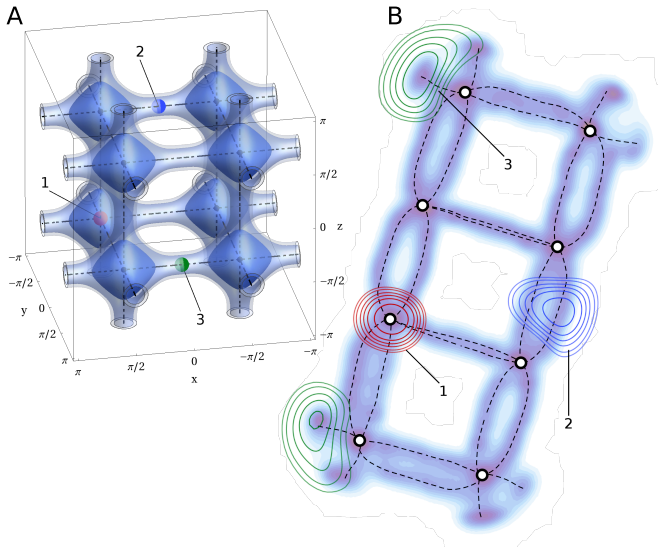

Fig. 1. A complex free energy surface that is periodic in three directions (A) and its sketch-map projection (B). Panel B shows how one can use functions of the sketch-map coordinates to describe the position in the three dimensional space (see methods section). The fields generated for the three marked points are shown. Where sketch-map reproduces the topology (point 1) the field is sharply peaked and is roughly Gaussian shaped. Where sketch-map provides less good description the field has multiple peaks because there are multiple points where it is reasonable to project (point 3).

Sketch-map produces a low-dimensional map of phase space in which the various basins in the high-dimensionality free-energy surface are well separated (16). As such sketch-map coordinates satisfy one of the conditions we require for a good collective variable, and free energy surfaces projected as a function of them are highly revealing. Where they fall short somewhat is in their description of the transition pathways between basins. This is to a certain extent unavoidable - representing complex features in a lower dimensionality space introduces distortions, which inevitably concentrate in poorly sampled regions such as the transition states. To clarify this issue a potential pitfall is illustrated in figure 1. This figure shows a three dimensional potential energy surface with periodic boundary conditions that contains eight energetic basins and twenty four transition pathways In projecting this landscape we must map a three-dimensional, toroidal space into a two dimensional plane. It is impossible to do this without introducing distortions much as it is impossible to map the surface of the earth on a flat surface rather than on the surface of a globe. Figure 1 also shows the two-dimensional representation of the surface generated by sketch-map. This projection is nevertheless revealing as it nicely separates the basins while mapping out most of the transition pathways in this free energy surface. Obviously though the mapping is imperfect - four of the transition pathways are distorted to the extent that points which are adjacent in the 3D representation are projected at opposite ends of the 2D representation. Consequentially, certain portions of the high-dimensionality space are not mapped out properly and will present a problem when this projection is used inside an enhanced sampling

algorithm. As we will discuss in the next section we have remedied this problem by developing a new framework for enhanced sampling, which exploits more of the information we obtain when we perform projections from the D-dimensional to the d-dimensional space.

Enhanced sampling algorithm

To enhance the sampling along the sketch map coordinates using metadynamics we must be able to calculate the projection (x) of any arbitrary point (X) in the D-dimensional space. Using a set of N landmark points X_i and their projections x_i one could compute a weight for each landmark based on the distance $|X - X_i|$ and then compute x as a weighted average. This is the basis of path collective variables (30) and a recently proposed method based on Isomap (31). It assumes that the $X_i s$ represent a dense sampling of of the high-dimensionality manifold and that the manifold is quasi-linear in the neighborhood of each landmark point. These assumptions are not valid for sketch-map coordinates, which endeavor to describe poorly sampled, highly-non-euclidean space. Hence, as discussed in our previous paper (16), a better approach for finding out-of-sample projections is to minimize the stress function:

$$\chi^2(X,x) = \frac{\sum_{i=1}^{N} w_i \left[F\left(|X - X_i|_D\right) - f\left(|x - x_i|_d\right)\right]^2}{\sum_{i=1}^{N} w_i}, \quad [1]$$

where w_i is the weight of the ith landmark point, and $F(x)$ and $f(x)$ are the sigmoid functions that were used to construct the sketch-map projection. Minimizing this function is problematic because in the vicinity of transition states where the sketch-map projection is poor there may be multiple nearly degenerate minima in the above. Consequentially, the d-dimensional projection of a trajectory in the D-dimensional space contains discontinuities and poor descriptions of some of the conformational transitions. We thus require an alternative approach that incorporates a better description of these problematic regions and in which any discontinuities are smoothed out. In our solution to this problem each high-dimensional configuration X is associated with a d-dimensional *field*, $\phi_X(x)$, which is given by:

$$\phi_X(x) = \frac{\exp\left(-\frac{\chi^2(X,x)}{2\sigma^2}\right)}{\int \exp\left(-\frac{\chi^2(X,x)}{2\sigma^2}\right) \mathrm{d}x} \quad [2]$$

This field replaces the usual representation based on d-dimensional points x. The overlap between fields, which measures their similarity, replaces the distance. The smearing parameter, σ, can be set by ensuring that the overlap between fields corresponding to structurally distinct landmark points is negligible. Then, with this in place, we can create an algorithm that is analogous to metadynamics (32) and use a history dependent bias to discourage the system from returning to previously-visited configurations. Now though this bias [1] is calculated from the overlap between the instantaneous field, $\phi_X(x)$, and a *bias field* $v(x,t)$, constructed from previously visited configurations:

$$V(X,t) = \int \phi_X(x) v(x,t)\, \mathrm{d}x, \quad [3]$$

where

$$v(x,t) = \sum_{t'=0}^{t} \omega \exp\left[-\frac{V(X(t'),t')}{\Delta T}\right] \phi_{X(t')}(x). \quad [4]$$

When χ^2 has a well-defined global minimum, this field is strongly concentrated about the minimum, with a shape that is nearly-Gaussian. Consequentially, the algorithm described above reduces to well-tempered metadynamics (33) in this limit (see supporting information). The pleasing thing though is that, as shown in figure 1, when the minimization is not straightforward the probability distribution splits itself between the various degenerate minima in equation 1. Therefore, these fields give a better description of the trajectory in regions where the sketch-map projection is poor. Furthermore, the field changes smoothly even when the out-of-sample projection changes discontinuously. A slight problem is that there is no longer a simple mathematical relationship between the final bias and the free energy surface. However, this is easily resolved as one can always reconstruct the free energy surface using on-the-fly re-weighting (8–11). In fact, calculating the free energy in this way is advantageous as the converged FES will not be affected if the fields are broader than the features in the free energy landscape. Hence, a poorly chosen σ will not adversely affect the accuracy of the method.

Results

Model potential. To test our new algorithm we first examined the model potential shown in figure 1. As we have explained here and in our previous paper (16), it is difficult to produce a two-dimensional, geometry-preserving map of the low energy portions of this potential. The sketch-map projection nicely separates the eight basins but only by introducing severe distortions in four of the transition pathways. These four discontinuities make the machinery discussed above absolutely

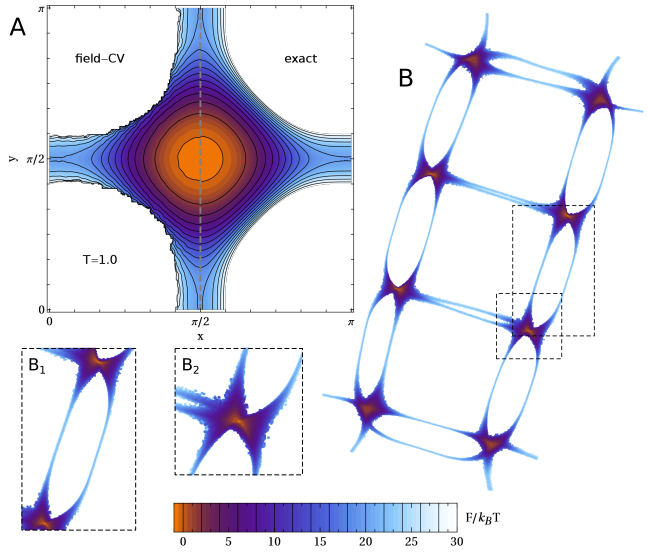

Fig. 2. Free energy surfaces for the model potential shown in figure 1 calculated by re-weighting the trajectories obtained from the field-overlap metadynamics simulations. Panel A shows the free energy as a function of the modulus of two of the three degrees of freedom. In this panel we compare the free energies obtained by re-weighting the trajectory with those calculated by explicitly integrating the free energy using the known Hamiltonian. Panel B shows the free energy surface as a function of the sketch-map coordinates. This was calculated by re-weighting the metadynamics trajectories and using the out-of-sample extension from reference (16) to define the instantaneous position in sketch-map space. Insets showing the free energy in the vicinity of one of the basins and along a pair of transition pathways are also shown.

[1] The corresponding force is equal to:

$$-\frac{\partial V(X,t)}{\partial X} = \frac{1}{2\sigma^2} \int \mathrm{d}x \phi_X(x) \left[v(x,t) - V(X,t)\right] \frac{\partial \chi^2(X,x)}{\partial X}$$

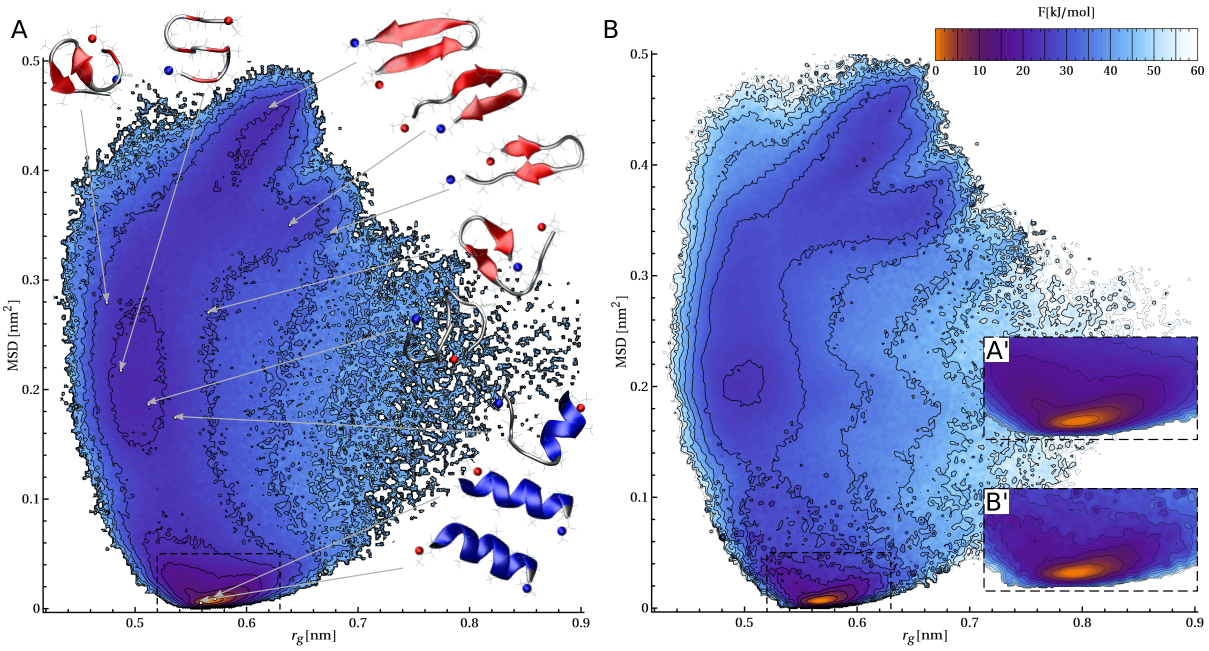

Fig. 3. The free energy landscape for ala12 in implicit solvent calculated from parallel tempering (panel A) and field-overlap metadynamics (panel B, using sketch-map coordinates) simulations. Here the free energy is shown as a function of the radius of gyration and the mean square displacement from the native state configuration. These CVs project radically different structures close together and thus many of the energetic barriers between structures disappear. In the highlighted structures the red and blue balls indicate the positions of the N and C termini respectively.

critical. The results shown in figure 2 show that our new algorithm performs admirably. We are able to quickly explore the entirety of the space, we see many re-crossing events and the bias converges by the end of our simulations (see the supporting information). Together these factors make it so that we can safely extract the free energies surfaces shown in figure 2 by re-weighting the histogram of visited configurations. In figure 2 we compare the re-weighted free energies with those obtained by integrating out explicitly one of the three degrees of freedom in figure 2. Also shown is the free-energy computed as a function of the sketch-map coordinates. This is perhaps more revealing as in this representation the complex topology of the free-energy surface with its eight identical basins can be clearly seen. This representation also demonstrates that there are six escape routes from each basin and that every pathway that is not broken by the projection (i.e. every pathway that we can examine using these CVs) is energetically equivalent.

Polyalanine-12. Having demonstrated our algorithm on a relatively simple energy landscape we now turn our attention to a more complex system; namely, the landscape of polyalanine-12 in implicit solvent. This system has been extensively studied and it has been shown that the potential energy surface, although very rough, is overall funnel-shaped with an alpha-helical global minimum (5,34). However, in spite of this structure, local minima in the potential energy surface (35) prevent the system from forming the helix during long, unbiased MD simulations (15), which suggests that MD alone is not a suitable tool for exploring this landscape. In contrast, reconnaissance metadynamics can find the global minimum so we have thus used this technique to collect the data (15) we used to construct sketch-map projections (16).

To asses the quality of our metadynamics simulations we first computed free energy surfaces for ala12 using parallel tempering (36, 37). In and of themselves these results are interesting as they demonstrate that, when free energies are displayed as a function of a set of simple collective coordinates, the resulting pictures can give a myopic view of the underlying physics. Figure 3 shows the free energy surface as a function of the gyration radius and the mean square displacement from an alpha helix configuration. This surface is very smooth, and has just two prominent features - one peaked basin and one very broad basin - which can be associated with the folded and unfolded states. This smoothness is in sharp contrast to the picture in figure 4, which shows the free-energy in terms of the sketch-map coordinates. In this representation the free energy surface appears to be very rough with a large number of very well-localized basins. This is more in line with what one would expect given the results from potential energy methods (34) and given that the system does not fold during a 1 μs MD simulation (15). Nevertheless, the CVs used in figure 3 can differentiate between the folded and unfolded states. Hence, they can safely be used to compare the relative populations of these states in unbiased MD simulations so that comparisons can be made with experiment. The problem though is that the description of the free energy landscape that these CVs provide is incomplete - these simple collective coordinates can distinguish folded configurations from the sea of unfolded states but are unable to detect the sometimes marked differences between the various unfolded configurations.

When CVs do not discriminate well between states interesting features in the landscape get blurred out. This fact explains why the many basins, which are visible in figure 4, become blurred into a broad, featureless valley in which there are no well-defined minima in figure 3. A direct consequence of this blurring is that, when these simple collective coordinates are used in a metadynamics simulation, the estimate of the free energy will converge very slowly. In fact, in all probability, it will only be possible to converge the free energy by combining metadynamics with parallel tempering (38) so as to ensure that barriers in the transverse degrees of freedom

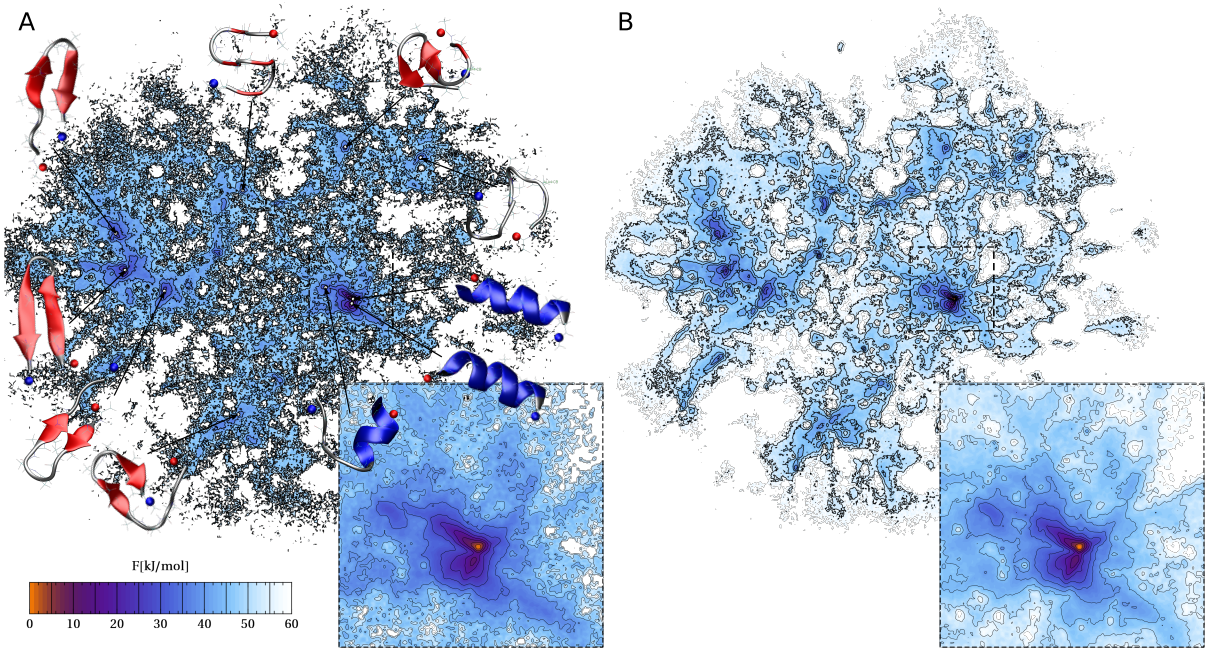

Fig. 4. The free energy landscape for ala12 in implicit solvent calculated from parallel tempering (panel A) and field-overlap metadynamics (panel B) simulations. Here the free energy is shown as a function of the sketch-map coordinates and is seen to be very rough. In contrast to figure 3 each of the highlighted structures lies in a separate basin in the free energy surface. In these structures the red and blue balls indicate the positions of the N and C termini respectively.

can be crossed. This is an expensive solution to the problem, however, that increases in expense with system size. Furthermore, one is left feeling that, were the collective coordinates better able to describe the various barriers to motion, much of this computational expense could be avoided.

In figures 3 and 4 we show the free energy surfaces obtained through on-the-fly re-weighting of field-overlap metadynamics simulations performed using sketch-map coordinates in tandem with the field formalism laid out previously. These free energy surfaces were calculated at 525 K, which is the unfolding temperature for this system. Reproducing the free energy surface at this temperature is particularly challenging because, unlike at 300 K, there is significant occupancy of the unfolded state. Even so the surfaces obtained from metadynamics are very close to the results from the parallel tempering (see supporting information for a more quantitative analysis). This result is particularly impressive given the complexity of the free-energy surface when it is projected as a function of the sketch-map coordinates (figure 4).

Conclusions

Enhanced sampling and free energy methods have been used to understand a wide variety of chemical and physical phenomena. In producing these successful applications, choosing collective variables that describe the problem of interest is critical. Making this choice is enormously difficult and it is perhaps this problem, more than any other, which prevents these methods being used even more widely. This choice will always be challenging, however, as we perform simulations to understand the properties of complex Hamiltonians that describe enormous numbers of interrelated energetic constraints. Given this we should not be surprised when the results cannot be explained using a simple function of the atomic positions.

When using collective variables in any approach, it is critical to remember that low-dimensional descriptions of complex chemical processes are inherently limited. Invariably certain features of the intrinsically high-dimensional process will be left out. This means that any decision as to what CV to use should be based on what information can be safely discarded. For example, if one wants to examine the folding equilibrium by analyzing an unbiased MD simulation, the CVs used in figure 3 are sufficient as these coordinates ably distinguish between folded and unfolded configurations. In contrast, where more detailed descriptions of the unfolded landscape are required these CVs fail because they cannot describe the subtleties in regions of phase space that are not in the immediate vicinity of the minimum-energy, folded state.

In biased MD choosing CVs is particularly critical as, for these methods, barriers to motion in orthogonal degrees of freedom can prevent free energy estimates from converging. Furthermore, given that in many applications one is endeavoring to accelerate rare events a thousand-fold or even a million-fold times, even barriers that are small compared to that of the rare event represent significant hurdles. Consequentially, using CVs that, like those used in figure 3, only distinguish the folded state from the sea of unfolded configurations in algorithms such as umbrella sampling or metadynamics will always be problematic. In these cases sketch-map CVs are a better approach as the data-driven strategy used to derive these coordinates ensures that distinct energetic basins are mapped to different parts of the low-dimensional space. The downside of this is that the sketch-map representation can contain discontinuities. However, as we have shown herein, this problem can be resolved by using fields to describe the instantaneous state. Admittedly, calculating the overlap integrals in this approach is considerably more computationally expensive than calculating the value of a CV. However, it is straightforward to parallelize these calculations on cheap GPU processors and, more importantly, unlike parallel tempering, the cost of this method is independent of system size. Hence, it can be used to calculate the free energies in very large systems or in *ab initio* calculations, where multiple replicas are less feasible. Furthermore, because the free energy is extracted

by re-weighting, it can be calculated as a function of any collective coordinate or examined using the collective variable free approaches that have been applied to the analysis of unbiased MD trajectories (27, 39, 40).

Dimensionality reduction is a generic tool that is used in fields of science ranging from chemistry and physics to social sciences and psychology. In all these fields this technique serves to identify low-dimensional trends in easy-to-measure, high-dimensionality data so that diverse features in the underlying phenomena can be classified. This understanding can then be used to classify points from outside the fitting set so that their likely behavior can be inferred. If this is done by minimizing equation 1, one is forced to assume that the fitting set describes every possibility and that the low-dimensional representation is a sensible topological description of the high-dimensionality data. In contrast, representing a configuration by a field like that in equation 2, allows one to perform these out-of-sample classifications more tentatively and to identify regions of the high-dimensional space where the low-dimensional representation is perhaps lacking. This approach is generic and builds on the notion that the overlap between normalized fields gives a measure of their similarity. In some cases – where it is natural to represent the high-dimensionality data using a normalized histogram (41) – it may even be possible to use the overlap between these probability distributions directly, and to avoid the dimensionality reduction step completely.

Materials and Methods

Re-weighting. All the free energy surface obtained from metadynamics simulations were calculated using on-the-fly re-weighting of multiple trajectories. The free energies as a function of a collective coordinate, s, were calculated based on a single trajectory using:

$$F(s)=-k_B T \log\left[\frac{\sum_{t'=1}^{t} \delta(s(t')-s)\exp\left(+\frac{V(X(t'),t')}{k_B T}\right)}{\sum_{t'=1}^{t}\exp\left(+\frac{V(X(t'),t')}{k_B T}\right)}\right] \qquad [5]$$

where the sum runs over the entirety of the trajectory. The free energies shown in the paper were then calculated by averaging the free energies obtained from a number of statistically uncorrelated simulations.

Model potential. The model potential shown in figure 1 is given by $V(\theta,\phi,\psi)=\exp[3(3-\sin^4(\theta)-\sin^4(\phi)-\sin^4(\psi))]-1$. We study the thermodynamics of a particle of mass m at temperature T. Hence, if one defines the unit of length as l^*, then the characteristic time unit, t^*, is equal to $\sqrt{\frac{m}{k_B T}}l^*$. To integrate the equation of motion we used the velocity Verlet algorithm with a timestep of 0.01 t^*. Temperature was kept fixed using a Langevin thermostat that had a relaxation time of 0.1 t^*. The sketch map projection of this landscape that was described in Ref. (16) was used throughout. The integrals in equation 3 and the equations for the forces were evaluated numerically on a 250 $\times$ 250 grid of points. However, because evaluating the value of equation 1 at every one of these points would be prohibitively expensive, we chose instead to only evaluate this function on a 15 $\times$ 15 grid of points. The function was then interpolated onto the remaining grid points using a bicubic interpolation algorithm (42). The bias field was augmented with a new function every 100 steps, while the initial height, ω, and the well tempered factor, ΔT, were set equal to 0.44 $k_B T$ and 4 $k_B T$ respectively. To collect adequate statistics the free energy surfaces shown in figure 2 were calculated from sixteen statistically-uncorrelated runs, which each ran for a total time of 52800 t^*.

Alanine 12. All simulations of polyalanine were run using gromacs-4.5.1 (43), the amber96 forcefield (44) and a distance dependent dielectric (34). A time step of 1 fs was used throughout, all bonds were kept rigid using the LINCS algorithm, and the van der Waals and electrostatic interactions were calculated without any cutoff. The temperature was maintained using an optimal-sampling, colored noise thermostat (45). Once again the sketch-map projection from Ref. (16) was used and integrals were calculated on a 401 $\times$ 401 grid of points that was constructed by performing a bicubic interpolation from a sparser 21 $\times$ 21 grid of points. The bias field was augmented with a new function every 500 steps, while the initial height, ω, and the well tempered factor, ΔT, were set equal to 0.5 kJ mol^{-1} and 2100 K respectively. The free energy surfaces shown in figures 3 and 4 were calculated by re-weighting from 16 such runs - a total of 800 ns of simulation time. For comparison we also calculated the free energy surface for this system from a single, 800 ns parallel tempering calculation with 5 replicas in which swapping moves were attempted every 100 steps. The temperatures of the replicas in this calculation were 525.00 K, 601.86 K, 688.23 K, 785.17 K and 886.17 K. The radius of gyration and distance from the alpha-helical configuration were calculated using PLUMED (13).

ACKNOWLEDGMENTS. The authors would like to thank Davide Branduardi and Giovanni Bussi for useful discussions along with Ali Hassanali and Federico Giberti for reading early drafts of the manuscript and giving suggestions. This work was funded by the European Union (Grant ERC-2009-AdG-247075), the Royal Society, and the Swiss National Science Foundation.

1. Garcia, AE (1992) Large-amplitude nonlinear motions in proteins. Phys. Rev. Lett. 68:2696–2699.
2. Amadei, A, Linssen, ABM, Berendsen, HJC (1993) Essential dynamics of proteins. PROTEINS: struct. funct. gen. 17:412.
3. Piana, S, Laio, A (2008) Advillin folding takes place on a hypersurface of small dimensionality. Physical Review Letters 101:208101.
4. Hegger, R, Altis, A, Nguyen, PH, Stock, G (2007) How complex is the dynamics of peptide folding? Phys. Rev. Lett. 98:028102.
5. Wales, DJ (2003) Energy Landscapes (Cambridge University Press).
6. Shaw, DE et al. (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330:341–346.
7. Frenkel, D, Smit, B (2002) Understanding Molecular Simulation (Academic Press).
8. Kumar, S, Bouzida, D, Swendsen, RH, Kollman, PA, Rosenberg, JM (1992) The weighted histogram analysis method for free-energy calculations on biomolecules i the method. Journal of Compuational Chemistry 13:1011–1012.
9. Dickson, BM, Lelièvre, T, Stoltz, G, Legoll, F, Fleurat-Lessard, P (2010) Free energy calculations: An efficient adaptive biasing potential method. J. Phys. Chem. B 114:5823.
10. Bonomi, M, Barducci, A, Parrinello, M (2009) Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics. J. Comput. Chem. 30:1615.
11. Dickson, BM (2011) Approaching a parameter-free metadynamics. Phys. Rev. E 84:037701.
12. Laio, A, Gervasio, FL (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and materials sciences. Reports on Progress in Physics 71:126601.
13. Bonomi, M et al. (2009) Plumed: A portable plugin for free-energy calculations with molecular dynamics. Computer Physics Communications 180:1961 – 1972.
14. Barducci, A, Bonomi, M, Parrinello, M (2011) Metadynamics. Wiley Interdisciplinary Reviews: Computational Molecular Science 1:826–843.
15. Tribello, GA, Ceriotti, M, Parrinello, M (2010) A self-learning algorithm for biased molecular dynamics. Proceedings of the National Academy of Sciences 107:17509–17514.
16. Ceriotti, M, Tribello, GA, Parrinello, M (2011) Simplifying the representation of complex free-energy landscapes using sketch-map. Proceedings of the National Academy of Sciences.
17. Geissler, PL, Dellago, C, Chandler, D (1999) Kinetic pathways of ion pair dissociation in water. The Journal of Physical Chemistry B 103:3706–3710.
18. Cox, TF, Cox, MAA (1994) Multidimensional Scaling (London: Chapman and Hall).
19. Tenenbaum, JB, Silva, Vd, Langford, JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323.
20. Das, P, Moll, M, Stamati, H, Kavraki, LE, Clementi, C (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proceedings of the National Academy of Sciences 103:9885–9890.
21. Schölkopf, B, Smola, A, Müller, KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Compuation 10:1299–1319.
22. Schölkopf, B, Smola, A, Muller, KR (1999) in Advances in Kernel Methods-Support Vector Learning (MIT Press), pp 327–352.
23. Coifman, RR et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. Proceedings of the National Academy of Sciences of the United States of America 102:7432–7437.
24. Coifman, RR, Lafon, S (2006) Diffusion maps. Applied and Computational Harmonic Analysis 21:5 – 30 Diffusion Maps and Wavelets.
25. Belkin, M, Niyogi, P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15:1373–1396.
26. Ferguson, AL, Panagiotopoulos, AZ, Debenedetti, PG, Kevrekidis, IG (2010) Systematic determination of order parameters for chain dynamics using diffusion maps. Proceedings of the National Academy of Sciences 107:13597–13602.

27. Rohrdanz, MA, Zheng, W, Maggioni, M, Clementi, C (2011) Determination of reaction coordinates via locally scaled diffusion map. The Journal of Chemical Physics 134:124116.
28. Donoho, DL, Grimes, C (2002) When does isomap recover the natural parameterization of familes of articulated images?, (Department of Statistics, Stanford University), Technical Report 2002-27.
29. Donoho, DL, Grimes, C (2003) Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences of the United States of America 100:5591–5596.
30. Branduardi, D, Gervasio, FL, Parrinello, M (2007) From a to b in free energy space. J. Chem. Phys. 126:054103.
31. Spiwok, V, Kralova, B (2011) Metadynamics in the conformational space nonlinearly dimensionally reduced by isomap. The Journal of Chemical Physics 135:224504.
32. Laio, A, Parrinello, M (2002) Escaping free energy minima. Proc. Natl. Acad Sci U.S.A. 99:12562.
33. Barducci, A, Bussi, G, Parrinello, M (2008) Well tempered metadynamics: A smoothly converging and tunable free energy method. Phys. Rev. Lett. 100:020603.
34. Mortenson, PN, Evans, DA, Wales, DJ (2002) Energy landscapes of model polyalanines. J. Chem. Phys. 117:1363.
35. Dill, KA, Ozkan, SB, Shell, MS, Weikl, TR (2008) The protein folding problem. Annual Review of Biophysics 37:289–316 PMID: 18573083.
36. Sugita, Y, Okamoto, Y (1999) Replica-exchange molecular dynamics for protein folding. Chem Phys Lett. 314:141.
37. Hansmann, UE (1997) Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett. 281:140.
38. Bussi, G, Gervasio, FL, Laio, A, Parrinello, M (2006) Free-energy landscape for hairpin folding from combined parallel tempering and metadynamics. Journal of the American Chemical Society 128:13435–13441 PMID: 17031956.
39. Krivov, SV, Karplus, M (2002) Free energy disconnectivity graphs: Application to peptide models. Journal of Chemical Physics 117:10894–10903.
40. Gfeller, D, De Los Rios, P, Caflisch, A, Rao, F (2007) Complex network analysis of free-energy landscapes. Proceedings of the National Academy of Sciences 104:1817–1822.
41. Tribello, GA, Cuny, J, Eshet, H, Parrinello, M (2011) Exploring the free energy surfaces of clusters using reconnaissance metadynamics. Journal of Chemical Physics 135:114109.
42. Press, WH, Teukolosky, SA, Vetterling, WT, Flannery, BP (2007) Numerical Recipes: The art of scientific computing (Cambridge University Press).
43. Hess, B, Kutzner, C, van der Spoel, D, Lindahl, E (2008) Gromacs 4: Algorithms for highly efficient, load-balanced and scalable molecular simulation. J. Chem. Theory Comput. 4:435.
44. Kollman, PA (1996) Advances and continuing challeges in achieving realistic and predictive simulations of the properties of organic and biological molecules. Acc. of Chem. Res. 29:461.
45. Ceriotti, M, Bussi, G, Parrinello, M (2010) Colored-noise thermostats la carte. Journal of Chemical Theory and Computation 6:1170–1180.